PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits

Year: 2023 | *Citations: 181* | *Authors: Hang Jiang, Xiajie Zhang, Xubo Cao*

Abstract

Despite the many use cases for large language models (LLMs) in creating personalized chatbots, there has been limited research on evaluating the extent to which the behaviors of personalized LLMs accurately and consistently reflect specific personality traits. We consider studying the behavior of LLM-based agents which we refer to as LLM personas and present a case study with GPT-3.5 and GPT-4 to investigate whether LLMs can generate content that aligns with their assigned personality profiles. To this end, we simulate distinct LLM personas based on the Big Five personality model, have them complete the 44-item Big Five Inventory (BFI) personality test and a story writing task, and then assess their essays with automatic and human evaluations. Results show that LLM personas' self-reported BFI scores are consistent with their designated personality types, with large effect sizes observed across five traits. Additionally, LLM personas' writings have emerging representative linguistic patterns for personality traits when compared with a human writing corpus. Furthermore, human evaluation shows that humans can perceive some personality traits with an accuracy of up to 80%. Interestingly, the accuracy drops significantly when the annotators were informed of AI authorship.